Taux de défaillance bayésien pondéré à partir de données de retour d'expérience.

Brice Lanternier *,— Dominique Charpentier*— Patrick Lyonnet ****

** INERIS*
Parc ALATA
60550 Verneuil en Halatte
France

{brice.lanternier-étudiant, dominique.charpentier}@ineris.fr

*** LTDS de l'UMR 5513*
42000 ENISE- St Etienne France

patrick.lyonnet@enise.fr

RESUME : *Afin de réaliser une étude de sécurité, il est désormais nécessaire d'évaluer de manière probabiliste la sûreté de fonctionnement des matériels. Les taux de défaillance indispensables pour ces études peuvent s'obtenir à partir des bases de données. L'intérêt de ces bases est indéniable mais leur utilisation est rendue difficile par les différences de valeurs trouvées dans celles-ci. Cet article propose une méthode pour améliorer la prise en compte des données issues de différentes sources. La démarche utilise les techniques bayésiennes avec un aspect innovant dans la pondération des données d'entrées notamment par la quantité d'information de Fischer.*

ABSTRACT : *From now on, it is necessary to evaluate probabilistically the materials 's operation safety in order to achieve a safety analyse. The essential failure rates for those kind of studies can be recovered thanks to reliability database. The interest is undeniable but the use is made difficult due to differences encountered in the data from one source to the other. This article propose a method to improve the taking into account of the data resulting from various sources. The study use Bayesian statistics with an innovating aspect in the entering data weighting in particular by introducing Fischer's information quantity.*

MOTS CLES : *Bayésien, Base de données de fiabilité, Quantité d'information de Fischer.*

KEYWORDS : *Bayesian, Reliability Data Bank, Fischer's information.*

1. Introduction

Pour calculer la fiabilité d'un matériel, il est indispensable d'avoir des informations sur les défaillances auxquelles il peut être soumis. Le Retour d'EXpérience (REX) fournie ces informations mais est contraignant à réaliser. Parmi les nombreuses exigences à satisfaire, peut être citer les temps de « retour de résultats » (nombre de défaillances représentatifs) qui nécessitent un investissement financier et organisationnel sur le long terme. Ce REX n'est par conséquent pas toujours présent dans les entreprises. A cela, plusieurs raisons ; il peut s'agir d'un matériel très fiable et pour lequel aucune défaillance n'a donc été constatée. Il peut également s'agir d'un matériel jamais utilisé auparavant en interne. Dans tous ces cas, la seule solution pour obtenir des données est l'utilisation de bases de données.

Il existe, en effet, un certain nombre de bases de données recensant différents matériels et, qui pour chacun d'eux, donnent un taux de défaillances.

Lorsque l'on désire utiliser ces bases, on décèle rapidement certaines difficultés quant à l'appropriation des données qui s'y trouvent. Le matériel peut ne pas avoir été analysé dans la base de données. De grosses différences de résultat s'observent au niveau des taux de défaillance annoncés pouvant varier de 1 à 100. Ces différences sont essentiellement dues à des caractéristiques différentes (conception, mode d'utilisation, maintenance…) qui ne transparaissent pas clairement entre les bases d'informations.

Nous avons développé une méthode d'utilisation optimisée de ces bases afin de conduire de façon plus rigoureuse une étude probabiliste de sûreté de fonctionnement. L'idée directrice est de tenir compte de l'ensemble des informations apportées par ces sources pour obtenir un taux de défaillance unique, intégrant de façon rigoureuse toutes les informations des données.

L'aspect innovant de cette approche est l'utilisation des statistiques bayésiennes (étant donné l'hétérogénéité des données à traiter) appliquée à des bases de données élaborées par différentes équipes à travers le monde dans des domaines d'activité variés. L'autre aspect innovant de cette démarche est de proposer un ensemble de pondération des sources mis en place notamment grâce à l'information de Fisher. Ainsi le modèle permet de quantifier l'information apportées par les bases de données et de pondérer l'apport de chacune. Une pondération des sources en fonction de critères évalués par l'analyste est également présente pour répondre à l'exigence de similarité entre ce qui a été analysé et ce que l'on désire étudier.

2. Modélisation

2.1. Hypothèse de départ

Cette étude repose sur les principes suivants:

— Les taux de défaillance sont constants. Cette première hypothèse se fonde sur le fait que nous utilisons des données existantes dans lesquelles la chronologie des défaillances n'est pas explicitée.

— La volonté de rendre compte d'un taux de défaillances prenant en considération le mode de fonctionnement nous a conduit à différencier les matériels fonctionnant à la sollicitation et les matériels fonctionnant en continu. Cependant, seul le fonctionnement en continu sera traité dans la suite de cet article.

Les temps de défaillances de matériels en fonctionnement suivent une loi exponentielle qui est directement associée au processus poissonien (représentatif de la durée de vie de matériels ayant un taux de défaillance aléatoire constant). Pour un fonctionnement en continu, la vraisemblance est donc modélisée par une distribution exponentielle.

$$\ell(k, T / \lambda) = \lambda^k e^{-\lambda T}$$

avec k représentant le nombre de défaillances, T le temps de bon fonctionnement cumulé et λ le taux de défaillances en continu.

2.2. Analyse Bayésienne

La mise en commun des données par approche Bayésienne a comme intérêt :

— d'utiliser le maximum d'informations disponibles,

— de valider un objectif de fiabilité avec un niveau de confiance,

— de permettre l'actualisation du savoir automatiquement,

— de se prononcer dans le cas où aucune défaillance n'a été constatée,

— de trouver une valeur (taux de défaillance) la plus pertinente.

Plusieurs cas de figures sont envisagés selon les données disponibles. Trois cas couvre le champ des possibilités :

— Une connaissance faible de l'a priori

— une connaissance moyenne de l'a priori

— une connaissance forte de l'a priori

2.2.1. Connaissance faible de l'a priori

Cette situation correspond soit à des avis d'expert peu précis soit à des bases de données donnant uniquement un taux de défaillance moyen. Dans le modèle proposé, cet état de la connaissance s'appuie sur les règles de Jeffreys (Lawless, 1987) et sera modélisé par une loi uniforme non informative.

$$g(\lambda) = \begin{cases} \frac{1}{\lambda_0}; \lambda_0 \in]0;\infty[\\ 0, \lambda_0 = 0 \end{cases}$$

La distribution a posteriori s'écrit donc :

$$g(\lambda / T) = \frac{T^{k+1} \cdot \lambda^k e^{-\lambda T}}{\Gamma(k+1, \lambda_0 T)}$$

L'espérance de cette distribution correspond à l'estimateur du taux de défaillance :

$$\hat{\lambda} = \frac{\Gamma(k+2, \lambda_0 T)}{T \cdot [\Gamma(k+1, \lambda_0 T)]}$$

2.2.2. Connaissance moyenne de l'a priori

Le déficit de connaissance ne nous permet ici de modéliser notre source que par une loi uniforme informative. On se trouve dans le cas d'un intervalle borné et toutes les valeurs possibles de la probabilité sont équiprobables sur cet intervalle.

$$g(\lambda) = \begin{cases} \frac{1}{\lambda_1 - \lambda_0}; \lambda \in [\lambda_0, \lambda_1] \\ 0, \lambda \notin [\lambda_0, \lambda_1] \end{cases}$$

Et finalement :

$$g(\lambda / T) = \frac{T^{k+1} \cdot \lambda^k e^{-\lambda T}}{\Gamma(k+1, \lambda_1 T) - \Gamma(k+1, \lambda_0 T)}$$

L'espérance de cette distribution correspond à l'estimateur du taux de défaillance :

$$\hat{\lambda} = \frac{\Gamma(k+2, \lambda_1 T) - \Gamma(k+2, \lambda_0 T)}{T \cdot [\Gamma(k+1, \lambda_1 T) - \Gamma(k+1, \lambda_0 T)]}$$

2.2.3. Connaissance forte de l'a priori

La modélisation retenue pour le taux de défaillance en continu est une distribution Gamma. Elle est en effet intéressante à utiliser dans notre cas puisqu'elle permet de représenter les trois phases de vie (courbe en baignoire) d'un système. L'hypothèse d'une vraisemblance de type exponentiel nous permet également, grâce à l'utilisation de la loi Gamma de pratiquer le principe de conjugaison d'une inférence bayésienne.

2.3. L'étape de pondération

2.3.1. Introduction

Une fois la modélisation concernant l'a priori et la vraisemblance réalisée, il reste à optimiser la pertinence de celles-ci. Il faut donc vérifier que la loi a priori déterminée à partir d'une ou de plusieurs sources n'influence pas de manière négative la vraisemblance et vice-versa, ce qui aurait pour conséquence de dégrader le résultat de l'a posteriori.

Cette nécessité de mettre en parallèle une distribution a priori, compatible avec le retour d'expérience, pour obtenir des résultats finaux réalistes d'une part, mais surtout homogène au système étudié, est pris en compte dans cette partie. Dans cette optique, le modèle proposé fait intervenir une méthodologie de pondération des sources qui permet de rationaliser a priori et vraisemblance.

Des études ont déjà proposé des solutions de pondérations permettant d'atténuer les effets d'une modélisation a priori erronée sur l'a posteriori, (Peruggia, 1997), (Ringler, 1988), (Usureau, 2001), (Yuille et *al*., 1996).

2.3.2. Pondération proposée

Trois critères distincts permettent la pondération à la fois de l'a priori et de la vraisemblance.

L'analyste est dans un premier temps, amené à juger de la similarité relative du système analysé dans les bases et celui qu'il désire étudié. Les aspects de conception, fabrication, process et milieu avec lesquels le matériel est en relation seront donc pris en compte.

Le deuxième critère permettant de pondérer les valeurs observées est la qualité de la source. Ce critère prend en compte la justesse relative de la source et correspond à la confiance de l'analyste sur la source. Il se basera donc sur la qualité des concepteurs de la base, sur la quantité et la manière de recueillir les renseignements, sur la cohérence des méthodes employées et de l'analyse.

Le troisième critère mis en place dans notre modèle prend en compte la quantité d'information apportée par chaque source. Ce critère s'appuie sur l'information mathématiquement rigoureuse de Fischer. L'information de Fischer permet en effet de s'adapter aux informations tant objectives que subjectives ainsi qu'aux lois usuelles de fiabilité. Par ailleurs, elle possède la propriété de l'additivité pour des sources indépendantes (Monfort, 1998). Cette quantité d'information dépend du nombre d'observations et de leur distribution. Elle est inversement proportionnelle à la variance de cette distribution. Elle reflète le volume de données contenues dans une source. Par cette pondération, on introduit la taille d'un échantillon complet « équivalent » à la taille de l'échantillon de la source de référence. Nous rappelons que l'information de Fischer notée $I(\lambda)$ est définie par :

$$I(\lambda \setminus x_i) = -kE\left[\frac{\partial^2 \ln L(x_i / \lambda)}{\partial \lambda^2}\right]$$

Les défaillances sont très informatives par rapport aux données censurées en C_i. Dans le cas de données fortement censurées, il est possible de prendre en compte la quantité d'information correspondante qui est alors :

$$I_c(\lambda \setminus x_i) = \frac{k(1 - e^{-\lambda \sum c_i})}{\lambda^2}$$

Cette quantité d'information permet de déterminer les coefficients de pondération à affecter aux échantillons statistiques afin que le poids relatif de chaque base soit (sauf avis contraire de l'analyste) la quantité d'information apportée par celles-ci relativement à la banque de référence. Les données sont alors directement comparables.

L'originalité de cette approche tient au fait que la pondération prend en compte les informations dont l'analyste dispose et donc des différentes situations auxquels il peut être confronté. A la différence des méthodes citées en 2.3.1, la pondération n'intervient pas uniquement sur l'a priori mais également sur la vraisemblance selon les situations rencontrées.

3. Application de l'utilisation de la quantité d'information de Fischer dans l'a priori

3.1. Présentation du problème

Les quatre bases de données de fiabilité (noté Bi) dans lesquelles des données plus ou moins informatives ont été extraites sont récapitulées dans le tableau 1avec k le nombre de défaillances et T le temps d'observation cumulé.

Sources	k	T(*E5)
B_1	2	0,69
B_2	7	1,96
B_3	102	221
B_0	148	31,8

Tableau 1. *Données provenant des bases Bi*

Des informations complémentaires sont évidemment présentes dans les différentes sources obtenues. L'analyste a estimé que l'adéquation au matériel était de 100% pour la base B_1, 50% pour B_2 et B_3 et 100% pour B_0. Ces estimations, qui interviendront dans l'estimation finale, ont été entreprises grâce à des jugements d'expert et par l'utilisation d'une approche semi paramétrique à taux proportionnels, qui ne sera pas développé dans cet article.

La distribution du taux de défaillances est par conséquent modélisée par une distribution gamma.

On calcul ensuite la quantité d'information apportée par chaque base (tableau 2).

Sources	λ (*E-6)	I (*E8)
B_1	29	23,8
B_2	35,7	54,9
B_3	4,61	47 883
B_0	46,5	683,3

Tableau 2. *Taux de défaillance et quantité d'information apportée*

Bilan:

$$I(B_1)/I(B_0)=1/29$$

$$I(B_2)/I(B_0)=1/12$$

$$I(B_3)/I(B_0)=70$$

On s'aperçoit que les données de B_1 et B_2 sont respectivement 29 et 12 fois moins informatives, les données de B_3 sont elles, 70 fois plus informatives que celles de B_0.

Après le choix d'une base de données comme référence, les sources vont être pondérées par la quantité d'information apportée. Cette étape ajoutée à la démarche Bayésienne classique permet de définir la taille d'un échantillon « équivalent » à l'échantillon de référence. On enrichit donc les données dans le cas où la source de référence est plus informative quantitativement que les autres et on procède à un nivellement dans le cas contraire. On détermine donc les coefficients de pondération à affecter aux échantillons statistiques afin d'égaliser (hypothèse de départ et sauf avis contraire de l'analyste) les quantités d'information apportées par chacune des banques relativement à la banque choisie comme référence. Les données sont alors directement comparables.

L'analyste peut souhaiter utiliser pour ses comparaisons les sources qu'il possède telles quelles : il y aura donc implicitement une pondération par la quantité d'information apportée par chacune des banques.

Nous considérons dans cet exemple que la qualité des quatre sources est identique. Ainsi nous ne prenons pas en compte ici ce critère afin de simplifier l'application.

Nous estimons que la source B_0 correspond le mieux aux matériels dont on désire évaluer le taux de défaillances. On fait alors le choix de pondérer toutes les sources par B_0.

Les données sont ensuite normalisées en définissant un nombre équivalent de défaillance et de temps de fonctionnement. On calcule alors les données fictives pondérées des banques autres que la référence.

Sources	k'	T'(*E5)
B_1	57,4	19,8
B_2	87,1	24,4
B_3	1,45	3,15

Tableau 3. *Défaillances et temps de fonctionnement normalisés*

Ces données vont constituer la fonction de vraisemblance dans une démarche bayésienne; la référence va constituer la distribution a priori qui sera jointe avec le même poids à la fonction de vraisemblance. L'a posteriori sera donc une « moyenne » équilibrée de toute l'information disponible.

Le choix de la source de référence ayant été effectué, la pondération par l'adéquation et la quantité d'information appliquée, on peut maintenant procéder à l'inférence bayésienne.

On obtient alors les paramètres de la distribution a posteriori ainsi que son espérance et un intervalle de crédibilité. Le taux de défaillance pertinent a posteriori obtenu est donc λ =38,7E-6 def/h.

3.2. Analyse

Sans pondération, il n'aurait pas été possible d'obtenir un taux de défaillance pertinent car la combinaison des différentes sources aurait été « faussée » par la base B_3.

En effet sans pondération le taux de défaillance obtenu serait λ=10,2 E-6 def/h (presque quatre fois moins) soit une estimation de la durée de vie augmenté de pratiquement 380%. Cette surestimation de la durée de vie peut être dangereuse dans le cas de dispositif de sécurité.

On s'aperçoit que sans pondération les données B_2 et surtout B_1 ont beaucoup moins de poids sur la vraisemblance à cause de leur faible informativité par rapport

à la source B_3. De plus, l'a priori perd un peu de son influence sur l'a posteriori par rapport à la vraisemblance. Au final, on remarque que la source B_3 a énormément d'influence sur l'a posteriori. En conséquence, on observe que dans le cas où la pondération n'a pas été effectuée, le taux de défaillance est très proche du taux de défaillance donné par la base ayant le plus de données. Celle-ci n'est pas forcément la plus juste et la plus pertinente par rapport au matériel étudié puisque les autres sources procurent des informations tout aussi intéressantes (voir adéquation du matériel).La pondération part les avis d'experts prend alors toute son importance.

Un échantillon fictif a posteriori représentatif de l'ensemble des informations contenues dans les bases de données et tenant compte de leurs incertitudes relatives (l'information de Fischer est fonction inverse de la variance) est déterminé. Cet échantillon peut constituer ultérieurement une fonction de vraisemblance si l'on veut créer une banque de données propre à une application. L'a priori sera alors constitué des données internes de cette industrie et des avis d'experts.

Cette démarche nécessite une bonne connaissance des outils mathématiques employés mais également des matériels étudiés afin de ne pas dégrader les résultats par rapport à ceux directement disponibles dans les Banques de Données de Fiabilité existantes. En effet, l'utilisation de la théorie Bayésienne ainsi que la mise en place de la pondération peut entraîner des aberrations s'ils sont utilisés avec négligence et sans analyse préalable.

4. Conclusion

Cette méthode permet de confronter plusieurs sources d'informations afin d'obtenir un taux de défaillance adéquat au matériel étudié. Pour obtenir ce taux de défaillance pertinent une pondération est mis en place. Elle s'appuie sur trois critères qui sont la qualité de l'information, l'adéquation au matériel et enfin la quantité d'information apportée qui est le critère prépondérant. Les résultats obtenus démontrent l'intérêt de cette méthode qui devrait jouer un rôle important dans les outils/méthodes basées sur les approches bayésiennes de la fiabilité.

5. Références

Lawless J.-F., « Statistical models and methods for lifetime data », John Wiley, 1987.

Monfort M.-L., « Estimation de la fiabilité d'un produit (nouveau ou existant) à améliorer à partir de retour d'expérience multiples et d'expertises », λμ 11, Arcachon, 1998.

Peruggia, « On variability of case-deletion importance sampling weights in the bayesian linear model », Congrès de Roumanie, 1997.

Ringler J., « Sur un petit problème pratique d'estimation en fiabilité », Revue de statistiques appliquées, 1988.

Usureau E., « application des méthodes bayésiennes pour l'optimisation des coûts de développement des produits nouveaux », thèse, ISTIA, 2001.

Yuille A.L., Bülthoff H.H., "Bayesian decision theory and psychophysics. Perception as Bayesian Inference", 123-161. (Eds.) Knill, D. and W. Richards, Cambridge University Press, Cambridge, 1996.